Past exam questions: Descriptive Statistics

1. A discrete numerical data set with $N$ numbers has the following relative frequency table:

<table>
<thead>
<tr>
<th>Relative frequency</th>
<th>0.25</th>
<th>0.1</th>
<th>0.2</th>
<th>0.2</th>
<th>$m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Points</td>
<td>-1</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

a) What is the relative frequency $m$?

$$m = 0.25$$

b) If there are $N = 50$ data points, what is the frequency of 0?

$$50 \times 0.1 = 5$$

c) What is the median? (Assume $N = 100$ data points).

$$\text{Median} = 1$$

d) What is the 40th percentile? (Assume $N = 100$ data points).

$$40^{th} \text{ percentile} = 1$$

2. The average of 18 numbers is 4, and the standard deviation is 1.2. What is the sum and the sum of squares of these numbers?

$$\bar{x} = \frac{\sum x}{n} = 4 \quad \rightarrow \quad \sum x = 4 \times 18 = 72$$

$$s = \sqrt{\frac{1}{n-1} \left[ \sum x^2 - \frac{(\sum x)^2}{n} \right]} = 1.2 \quad \rightarrow \quad \sum x^2 = 17 \times (4.2)^2 + \frac{(72)^2}{18} =$$

$$= 24.48 + 288 = 312.48$$
5) 20 data points, some sums and the correlation coefficient are given below.

<table>
<thead>
<tr>
<th>x</th>
<th>15</th>
<th>60</th>
<th>10</th>
<th>38</th>
<th>50</th>
<th>32</th>
<th>18</th>
<th>71</th>
<th>95</th>
<th>27</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>80</td>
<td>52</td>
<td>91</td>
<td>59</td>
<td>51</td>
<td>62</td>
<td>74</td>
<td>52</td>
<td>25</td>
<td>66</td>
</tr>
</tbody>
</table>

\[
\sum_{i=1}^{20} x_i = 956, \quad \sum_{i=1}^{20} y_i = 1102, \quad \sum_{i=1}^{20} x_i^2 = 60712, \quad \sum_{i=1}^{20} y_i^2 = 66592, \quad \sum_{i=1}^{20} x_i y_i = A, \quad r = -0.903.
\]

(a) What is the value of \( S_{xy} \)?

(b) What does \( r = -0.903 \) tell you? Explain.

(c) What is \( A \) (the sum of the cross products of the \( x_i \) and \( y_i \))? 

(d) Find the mean and standard deviation for the \( x_i \), and the mean for the \( y_i \).

(c) 

\[
S_{xy} = \frac{1}{20} \left( \sum_{i=1}^{20} x_i y_i - \frac{1}{20} \left( \sum_{i=1}^{20} x_i \right) \left( \sum_{i=1}^{20} y_i \right) \right) = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = -8478.89 \quad \left[ -8478.9, 8 \right]
\]

where 

\[
S_{xx} = 60712 - \frac{1}{20} (956)^2 = 15015.2
\]

\[
S_{yy} = 66592 - \frac{1}{20} (1102)^2 = 5871.8
\]

(b) The data is almost linear, with a negative slope.

(c) From (a), 

\[
S_{xy} = A - \frac{1}{20} (956)(1102) = -8478.89
\]

\[
\Rightarrow A = 44196.71 \quad \left[ 44201 \right]
\]

(d) 

\[
\bar{x} = \frac{956}{20} = 47.8, \quad \bar{y} = \frac{1102}{20} = 55.1
\]

\[
\text{Var}(x) = \frac{1}{19} S_{xx} = 250.76 \quad \Rightarrow \text{SD}(x) = 28.4
\]

(2)
6) The histogram of a set of discrete numerical data is shown below.

(a) Find the mode and the median.
(b) Find the sum of the data points. What is the sum of squares of the deviations?
(c) The 100p-th percentile is 2.5. What is p? (Give the answer in rational form)
(d) What is the interquartile range?
(e) What is the relative frequency of 2?

\[ \bar{x} = \frac{\sum x_i}{n} \]
\[ \sum (x_i - \bar{x})^2 = 111.78 \]

3) a) Mode = the most frequent = 4
b) Median = 50th percentile = \[ n \cdot p^{th} = 45(0.5) = 22.5^{th} \Rightarrow 23^{th} \]

4) sum = Mean \cdot 45 = 55
\[ \sum (x_i - \bar{x})^2 = 44(1.594)^2 \approx 111.78 \]

5 c.) 35th data point at 2 is 35th, first 3 is 36th.
\[ n = 35 \Rightarrow np = 35 \Rightarrow p = \frac{35}{45} \]

4 a.) \[ 45 \cdot \frac{1}{4} = 11.25 \Rightarrow 12 \]
\[ Q_1 = 0 \]
\[ Q_3 = 2 \]
\[ Q_3 - Q_1 = 2 \]

2 e.) \[ \frac{a}{45} \]
Q 1. (20 Points)

The histogram of a discrete data set consisting of 100 measurements is given below.

(a) Find the mode, median and mean.

(1) node: most frequent = 2

(2) median: \[ \frac{2 + 3.3}{2} = 2.65 \]

(2) mean: \[ \frac{\sum_{i=1}^{100} f(x) \cdot x}{100} = 3.115 \]

(b) Find the interquartile range.

\[ I.Q. \ range = Q_3 - Q_1 \]

(2) \[ Q_3 = \frac{3.3 + 5.2}{2} = 4.25 \]

(2) \[ Q_1 = 2 \]

(c) Find the frequency of 2.

(2) 35

(d) Find the sum of data points.

(3) \[ (15)(1.2) + (35)(2) + (25)(3.3) + (15)(5.2) + (10)(6.3) = 311.5 \]

(e) The variance is 2.687. What is \( \sum x_i^2 \)?

\[ s^2 = \frac{(\sum x_i^2 - (\frac{\sum x_i}{n})^2)}{n-1} = 2.687 = \frac{\sum x_i^2 - (311.5)^2}{100} \]

(2) \[ \Rightarrow \sum x_i^2 = 1236.336 \]

*Forgetting n-1: (2)
Q5. (15 P)
The IQ (Intelligence Quotient) scores of seven people are 86, 90, 94, 100, 108, 115, and 177.

(a) Find the mean and median for this data set.

\[ \bar{x} = \frac{\sum x_i}{n} = \frac{770}{7} = 110 \]

For the median, the data is already sorted in increasing order: 86, 90, 94, 100, 108, 115, 177 and 100 is the middle value; or

\[ n_p = 7 \times 0.5 = 3.5 \times 4 \text{th value} \]

so median is 100.

(b) Which is a better measure of center for this type of data? Why?

Median is a better measure of center, because the person with IQ score of 177 is an outliers (it is much larger than the rest of the scores). And mean is affected by outliers but not the median.

(c) Compare the range (sample range) and interquartile range for this data set.

\[ \text{Sample Range} = \text{largest} - \text{smallest} = 177 - 86 = 91 \]

\[ \text{IQR} = Q_3 - Q_1, \text{ where } Q_3 = 115 \text{ since } n_p = 7 \times 0.75 = 5.25 \text{ median value} \]

and \( Q_1 = 90 \), since \( n_p = 7 \times 0.25 = 1.75 \text{ 2nd value} \)

so \( \text{IQR} = 115 - 90 = 25 \).

(d) An IQ of 120 is said to be the 95th percentile. Explain what this means!

At least 95% of the IQ scores are at or below 120 and at least 5% of the IQ scores are at or above 120.
11. For a sample of 20 observations, the sum of the scores is 237 and the sum of squared scores is 3751. What is the value of the standard deviation? (3 points)

\[
\overline{x} = \frac{237}{20}, \quad \sum x^2 = 3751, \quad n = 20
\]

\[
\sigma^2 = \frac{1}{n-1} \left( \sum x^2 - \frac{\overline{x}^2}{n} \right) = \frac{1}{19} \left( 3751 - \frac{237^2}{20} \right) = 49.61
\]

\[
\Rightarrow \sigma = \sqrt{49.61} \approx 7.04
\]

12. What is the symbol associated with the value you calculated in #11? (2 points)

Use the following data set for problems 13-16:
35, 26, 18, 30, 2, 13, 29, 32, 24, 36, 35, 30

13. Find the mode(s) of this data set. (2 points)

30, 35

14. Find the range of this data set. (2 points)

36 - 2 = 34

15. Find the median of this data set. (3 points)

\[
\text{sorted}: 2, 7, 13, 18, 24, 26, 29, 30, 30, 32, 35, 35, 36
\]

\[n=12, \text{ median is the average of } 6^{th} \text{ and } 7^{th} \text{ values} = \frac{29+30}{2} = 29.5\]

16. Suppose a frequency table were constructed with class intervals of 1.5-9.5, 9.5-17.5, ..., 33.5-41.5 What is the relative frequency of observations for the fourth class interval? (3 points)
20. The saturated fat in grams ($X$) consumed per day for 16 males were: 
55, 65, 50, 34, 43, 58, 69, 36, 40, 95, 65, 70, 34, 35, 50
a. Calculate the mean. To receive full credit, give the symbol and the value. (3 points)
b. Find the mode and median. (4 points)
c. Which of the above three measures of center is most appropriate for this data and why? (4 points)
d. Calculate the standard deviation. To receive full credit, give the symbol and the value. (5 points)
e. Calculate the coefficient of variation (C.V.), and the IQR (interquartile range) for this data set. Would you use standard deviation, C.V., or IQR as a measure of spread for this data? Explain. (6 points)

Data sorted in ascending order: 4, 34, 35, 36, 39, 40, 43, 65, 50, 55, 58, 65, 65, 69, 70, 80, 95
(a) \( \bar{X} = \frac{\sum X}{n} = \frac{849}{16} = 53.0625 \)

(b) Two modes: 50, 65

Median: \( n = 16 \), so median is the average of 8th and 9th values.
Then median = \( \frac{50 + 55}{2} = 52.5 \)

(c) Median seems the best choice; since 4 can be viewed as an outlier.
But at the other extreme, there is another possible outlier, namely 95, so the data can also be viewed as symmetric. Hence, mean is also a reasonable choice.
Notice that both measures are very close to each other.

(d) \( s^2 = \frac{1}{n-1} \left( \sum X^2 - \frac{(\sum X)^2}{n} \right) = \frac{1}{15} \left( 52067 - \frac{849^2}{16} \right) \)

\( = \frac{1}{15} \left( 52067 - 45050.0625 \right) = \frac{7016.9375}{15} \)

\( = 467.796 \)

\( \text{std dev } s = \sqrt{467.796} = 21.63 \)

(e) C.V. = \( \frac{s}{X} \times 100 = \frac{21.63 \times 100}{53.0625} = 41.40 \% \)

IQR = \( Q_3 - Q_1 = 67 - 38 = 29 \)

IQR is the best choice here, since it is robust to outliers.
(17) Blood pressure values are often reported to the nearest 5 mmHg (100, 105, 110, etc.). Suppose the actual blood pressure values for nine randomly selected individuals are: 118.4, 127.4, 138.4, 130.0, 113.7, 122.0, 108.3, 131.5, 133.2.

(a) Calculate the mean and standard deviation for reported blood pressure of these individuals.

(b) What is the median of the reported pressure values?

(c) Suppose that the blood pressure of the second individual is 127.6 rather than 127.4. How does this affect the median of the reported value? What does this say about the sensitivity of the median to rounding or grouping in the data?

```
Reported blood pressure values in order:
110, 115, 120, 120, 125, 130, 130, 135, 140  n=9

a) \( \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \boxed{125} \)
   \( s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = (9.68)^2 \)
   \( s = \boxed{9.68} \)

b) \( n=9 \implies x_5 \) in the ordered list is the median
   \( \implies \) median is \( \boxed{125} \)

C) 127.6 is rounded to 130
   \( \implies \) new median is \( \boxed{130} \)

Median is highly sensitive to rounding in small data sets.
```
(21) There are three -3's, four -1's, two 0's, four 2's, and one 3 in a data set of temperatures (in degrees Celsius).

(a) Draw a relative frequency histogram
(b) Find the mode, median, mean, and 70th percentile.
(c) Repeat (b) after adding two -3's and two 3's to the temperature data set.
(d) Repeat (b) after converting the temperature data in Celsius to a temperature data in Fahrenheit.

(Degrees Fahrenheit = 1.8 • Degrees Celsius + 32. Example: 41°F = 5°C)

\[ \begin{align*}
&\text{Mode: } -1 \text{ and } 2 \\
&\text{Median: } -0.5 \\
&\text{Mean: } -0.14 = -\frac{1}{7} = -\frac{14}{14} \\
&70^{th} \text{ Percentile: } \frac{7}{10} \times 14 = 9.8 \rightarrow X_{10} = 2 \\
\end{align*} \]

\[ \begin{align*}
&\text{Mode: } -3 \\
&\text{Median: } -0.5 \\
&\text{Mean: } -\frac{2}{18} = -0.11 \\
&70^{th} \text{ Percentile: } \frac{7}{10} \times 18 = 12.6 \rightarrow X_{13} = 2 \\
\end{align*} \]

\[\begin{align*}
&\text{Mode: } -3 (1.8) + 32 = 26.6 \left[ \text{30.2 and 35.6} \right] \\
&\text{Median: } 31.1 \\
&\text{Mean: } 1.8(-0.11) + 32 \text{ or } 1.6(-0.14) + 32 \\
&\quad = 31.302 \quad = 31.74 \\
&70^{th} \text{ Percentile: } 1.8(2) + 32 = 35.6 \\
\end{align*} \]